REMARKS

Claims 1, 3-5, 7, 9, 11 and 13-15 have been amended. Claims 8, 12 and 16-17 have been canceled. Upon entry of the instant amendment, claims 1-7, 9-11, 13-15 and 18-20 will be pending in the application. Applicant respectfully requests reconsideration of the claims.

Claims 1-20 stand rejected under 35 USC 112, second paragraph, as being indefinite. Claim 1 has been amended to state that the goal of the method is to obtain and organize information about a plurality of unknown raw nucleic acid sequences. The information to be obtained and organized for each unknown raw nucleic acid sequence is the existence, descriptive information and qualifying properties of known nucleic acid sequences that are similar or identical to each unknown raw nucleic acid sequence.

With regard to claim 6, the subdirectories of claim 6 are created exactly where claim 6 states: "in each of the raw nucleic acid sequence directory, the trimmed nucleic acid sequence directory, the trimming parameters directory and the nucleic acid identification database search results directory." Regarding the "when" of the subdirectory creation, the subdirectories are created after the directories of claim 5 are created, when the user inputs a named individual library. See the specification at page 11, last paragraph.

Claim 7 has been amended to clarify when steps e) and f) are performed. Claim 12 has been canceled. In light of the amendments to the claims and the above discussion, applicant requests that the rejections of the claims under 35 USC 112 be withdrawn.

Claims 1-20 stand rejected under 35 USC 102(b) as being anticipated by Burland (2000). The instant application claims priority to provisional application 60/235,899 filed on September 28, 2000. Therefore, the rejection under 35 USC 102(b) is improper. At best, a rejection under 35 USC 102(a) may be possible. However, the date of the Burland document is 2000, with no month given. It is not clear that the Burland document was published prior to September 28, 2000, applicant's priority date. Should the examiner again reject any claim over Burland, applicant requests that the month and year of the publication of the Burland document be provided so that applicant may swear behind the Burland document, if necessary.

The examiner further states that the commercial product Lasergene, version 4.0, anticipates the claims. However, the examiner has not provided applicant a copy of the Lasergene product nor any evidence of its date as a reference. The mere fact that the product name is "Lasergene99" is not evidence that the product was actually available in 1999. Therefore, any rejections based on the product are not proper because applicant was not provided a copy of the product nor any evidence of its date as a reference.

Notwithstanding the above, the amended claims are distinguishable from Burland. Step c) of claim 1 has been amended to include additional features of the trimming step. The amendments to step c) of claim 1 are supported by, for example, pages 34-39 of the specification and the computer program contained on the compact disc submission. The other amendments to claim 1 are supported throughout the specification. No new matter has been added.

An important feature of claim 1 is the ability to manipulate **a plurality** of unknown raw nucleic acid sequences. Burland states that *SeqmanII* can assemble and trim several sequences into a **single contig** which is then submitted to NCBI's BLAST server. Burland, however, does not teach or suggest a method for submitting **a plurality** of sequences to a nucleic acid database, and obtaining search results for **each of the plurality** of sequences. In Burland, a separate manual request is needed for each search. That is, Burland does not teach automatic bulk processing of searches, as in claim 1.

Further, in Burland, the search results for a single sequence are exported from the html search results file by "cutting and pasting." In sharp contrast, the method of claim 1 **automatically creates** a plurality of first electronic spreadsheets (stored in computer files), with one first electronic spreadsheet containing the complete or a selected portion of the search results for each unknown raw sequence. Then, in step f) of claim 1, a second spreadsheet is automatically created using either computer determined best sequence matches or manual sequence selections from each of the first spreadsheets. The method of claim 1 is especially useful when the plurality of raw sequences are experimentally related. Then, all the information related to the experiment is concisely summarized in the second spreadsheet, with more detailed information available through hyperlinks to the first spreadsheets, the html files and the nucleic acid database. Using the automated method of claim 1, as opposed to manual cutting and pasting as in Burland, the applicant has noted up to a 90% reduction in human time required for unknown DNA identifier data capture.

Another important feature of claim 1 is step c), the trimming step. The inventive method automatically determines trimming locations of each unknown raw nucleic acid sequence using a scoring algorithm which, during each scoring operation, allows for nucleotide mismatches and either a single nucleotide insertion or a single nucleotide deletion in the unknown raw nucleic acid sequence. The scoring operations include matching a known 5' adapter sequence against the unknown raw nucleic acid sequence, matching a known 3' adapter sequence against the unknown raw nucleic acid sequence and matching a known confirmation sequence against the unknown raw nucleic acid sequence. Burland does not teach or suggest the detailed trimming feature recited in claim 1.

The inventive scoring algorithm first compares a known positive 5' adapter sequence and a known negative 5' adapter sequence (i.e., the 5' adapter sequence is compared in both a sense and antisense orientation) to the raw nucleic acid sequence and assigns a 5' trimming location to the position in the unknown raw nucleic acid sequence having the highest score. The method also determines the adapter insertion orientation according to the higher scoring of the known positive and negative 5' adapter sequences. The method then compares a known 3' adapter sequence having the same orientation as the higher scoring known 5' adapter sequence to the unknown raw nucleic acid sequence and assigns the 3' trimming location to the position in the unknown raw nucleic acid sequence having the highest score.

Then, a known positive confirmation sequence and a known negative confirmation sequence are compared to the raw nucleic acid sequence and a confirmation sequence location is assigned to a position in the unknown raw nucleic acid sequence having the highest score. The sequencing direction is determined according to the higher scoring of the known positive and negative confirmation sequences. This process is automatically performed for each of the plurality of sequences. Burland does not teach or suggest the claimed method of automatically determining trimming locations, as recited in step c) of claim 1 and discussed above.

Because of the many differences between claim 1 and Burland, claim 1 is allowable. The remaining claims depend directly or indirectly from claim 1 and are allowable for that reason. In addition, the dependent claims recite further features not found in Burland. For example, claim 7 allows steps d)-f) to be performed at a preset later time. Claims 9 and 11 recite features of the first and second spreadsheets. It should be noted that Burland nowhere discloses the concept of the concise summary provided by the second spreadsheet.

In light of the above, claims 1-7, 9-11, 13-15 and 18-20 are in condition for allowance. Should there be any questions regarding this application, the examiner is invited to contact the undersigned attorney at the number shown below.

Respectfully submitted,
William E. Eshelman
William E. Eshelman
Registration No. 35,865

Date: Oct. 14, 2003
William E. Eshelman
3130 Panhandle Road
Front Royal, VA 22630
(540) 636-6064

2000

5

DNASTAR's Lasergene Sequence Analysis Software

Timothy G. Burland

1. Introduction

Lasergene comprises eight applications, organized into functional units. A user with the full *Lasergene* system might employ the software as follows:

SeqManII: Trim and assemble sequence data, and determine the consensus sequence.
GeneQuest: Discover and annotate genes, patterns and other features among small, BAC-sized or larger sequences.
Protean: Predict protein secondary structure and identify antigenic regions.
MegAlign: Align sequences in pairwise or multiple configurations, and build phylogenetic trees.
GeneMan: Search sequence data with Boolean queries constructed from sequence similarity, consensus sequence and text terms.
PrimerSelect: Design primers for PCR, sequencing, hybridization, and transcription.
MapDraw: Create restriction maps displaying sites, translations, and features.
EditSeq: Import and manipulate sequences from other applications for analysis in *Lasergene.*

Examples of how the *Lasergene* applications may be used follow. The procedures are the same whether the software is run on *Windows 95/98/NT* or *Macintosh* computers. Examples are given for *Lasergene99* v4.0.

2. Sequence Assembly with *SeqManII*

SeqManII is the assembly software of choice for the *Escherichia coli* genome sequencing project *(1)*, among others. It can assemble sequencing projects

From: *Methods in Molecular Biology, vol. 132: Bioinformatics Methods and Protocols*
Edited by: S. Misener and S. A. Krawetz © Humana Press Inc., Totowa, NJ

ranging from less than a kilobase to as large as a bacterial genome. As an example, this section describes how almost 1200 ABI®-format sequence trace files from the *E. coli* genome project were assembled into one 93-kb contig.

2.1. Sequence Entry

The simplest way to add large numbers of sequence files to a *SeqManII* project is by drag-and-drop from the *Windows Explorer* or *Macintosh Finder.* Multiple folders may be added, and files may be ABI and SCF3 trace files, or *DNASTAR* sequence files. *SeqManII* accepts up to 64,000 sequences, ample for >5x shotgun sequence coverage of most bacterial genomes. Once sequences are added, a file of filenames (fof) can be created so that adding the same set of sequences to future assemblies can be done by adding the fof rather than the individual files.

2.2. Trimming Vector and Host Data

Sequence data may be contaminated with sequences from the host used to grow clones, and from the vector used to clone the target sequences. Such contaminants can compromise sequence assembly seriously. *SeqManII* removes contaminating vector or host sequences prior to assembly. To trim vector from the *E. coli* sequence reads, the sequences are sorted into two groups then entered into *SeqManII.* The forward reads are marked as cloned in the Janus vector *(2)* and the reverse reads are marked with the InvJanus vector from the vector catalog. For trimming, the recommended default stringency is used for sequence similarity to the vectors.

2.3. Quality Trimming

Sequence reads vary in quality over their length, with poor quality data typically occurring near 5′ and 3′ ends of each read. Poor-quality regions contain frequent basecalling errors, which compromise sequence assembly. *SeqManII* evaluates peak quality directly in fluorescent trace data, and trims data that fall below a specified quality threshold *(3)*. This retains good data and removes poor data without the need for human editing. In this *E. coli* example, quality trimming stringency was set to the recommended medium value.

After trimming, another fof may be saved, which stores the trimming information as well as the filenames. Thus, trimming need not be repeated for additional assemblies involving the same sequences.

2.4. Sequence Assembly

If all sequence reads originate from one contiguous piece of DNA, and there are reads for every segment of DNA, it should be possible to assemble the reads into one contiguous piece—or contig—corresponding to the original

ıe. As an
ɜnce trace
› contig.

SeqManII
ʼh Finder.
e files, or
es, ample
ɩequences
ɪme set of
r than the

st used to
ɔes. Such
ʼeqManII
. To trim
vo groups
ɩed in the
us vector
ngency is

data typi-
ɩs contain
ЅeqManII
a that fall
removes
e, quality

ing infor-
. for addi-

and there
ɜmble the
: original

piece of DNA. For sequence assembly, several parameters may be adjusted, including the extent of match needed to join sequence reads into contigs, and the penalties applied for introducing gaps in alignments of overlapping sequence reads. For most data, default values give the best assemblies—they provide the fewest number of contigs and minimize the possibility of false joins.

Once adjustable parameters are chosen, *SeqManII* can execute vector and quality trimming and sequence assembly with one click of the mouse. *SeqManII* trimmed and assembled the 1200 *E. coli* sequence reads into one 93-kb contig (**Fig. 1**) in 15–20 min on a 200-MHz PentiumPro™ Windows PC or Macintosh G3.

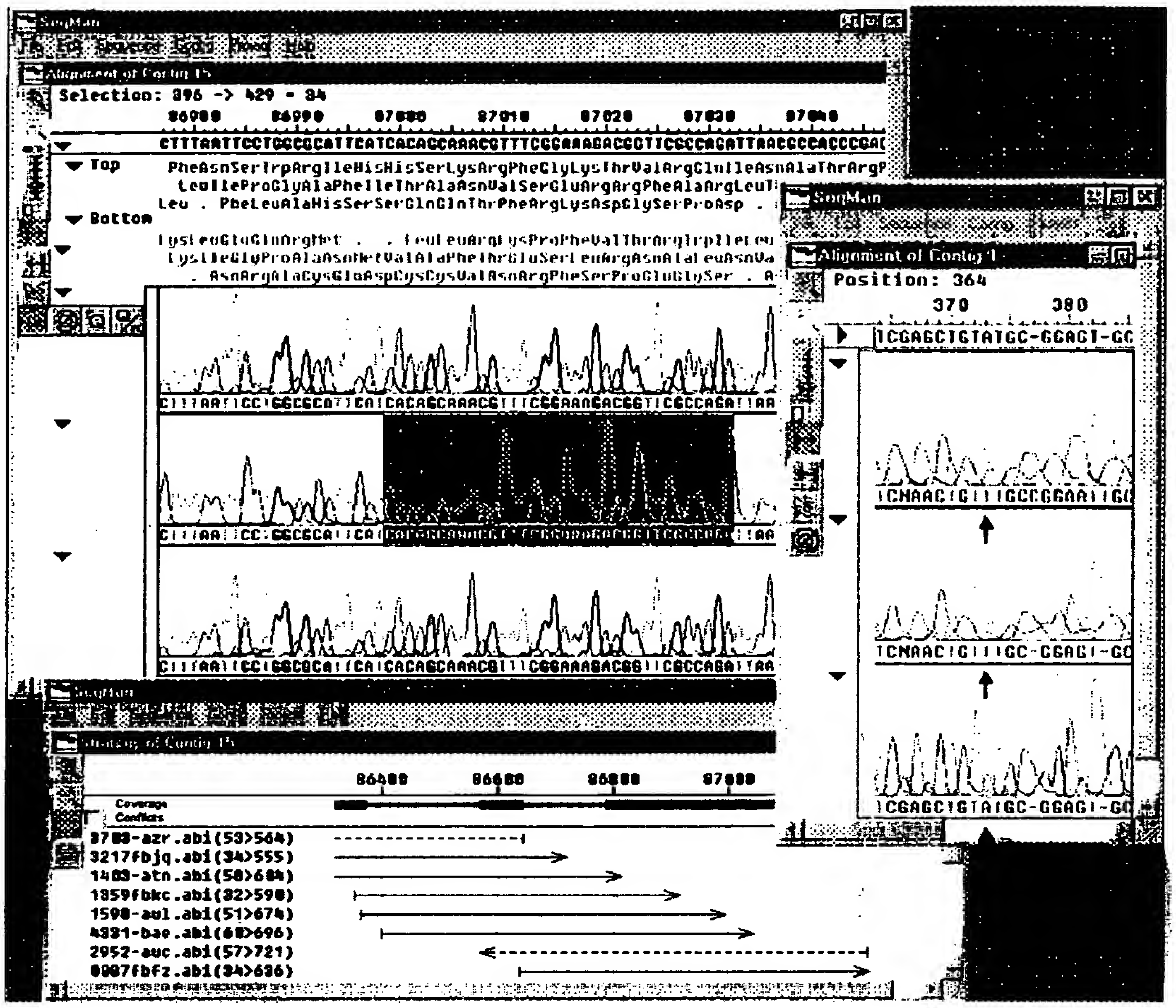
2.5. Consensus Calling

Once sequences are aligned and assembled (**Fig. 1**), the consensus sequence is called. *SeqManII,* like other assembly programs, can choose the base occurring most frequently at a given position for the consensus—a majority calling system. However, where fluorescence trace data are available, as in this project, *SeqManII* determines the consensus sequence based on evaluation of peak quality. This system is much more accurate than majority-based consensus calling *(3)*. Unlike majority-based systems, *SeqManII* can call the consensus correctly when a majority of the original basecalls are in error (**Fig. 1**). Such functionality is a prerequisite of assembly software suitable for the >99.99% accuracy goal of the human genome project.

Changes to the consensus calling criteria may be made without repeating the assembly. For trace data, the **Evidence Percentage** parameter controls the stringency used to make unambiguous consensus calls—i.e., specific base calls an International Union of Biochemistry (IUB) ambiguity code representing a heterozygous or questionable position. Setting **Evidence Percentage** to a higher value increases the likelihood that ambiguous or heterozygous calls will be made for the consensus. Setting **Evidence Percentage** too high might result in spurious heterozygous or ambiguous calls in the consensus. Conversely, reducing **Evidence Percentage** too far increases the risk of making an unambiguous call in the consensus when the evidence for that call is equivocal, or when the position may be heterozygous.

2.6. Editing

Although data trimming, assembly, and consensus calling may be run automatically, *SeqManII* provides a graphical interface for manual editing of the sequence reads and consensus (**Fig. 1**). Sequence reads, traces, and six-frame translation are all aligned with the consensus sequence. Open reading frames and stop codons in the translation direct attention to potential frameshifts or other sequencing problems.



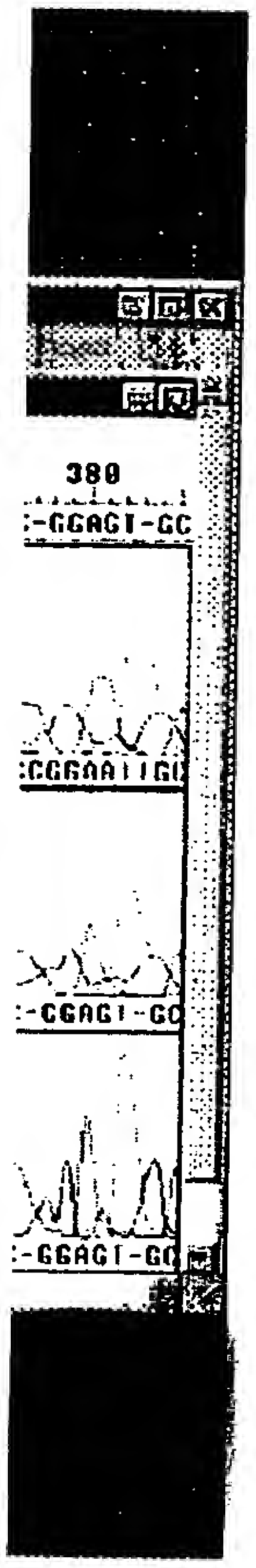
```
Selection: 396 -> 429 = 34
        86980     86990     87000     87010     87020     87030     87040
        CTTTAATTCCTGGCGCATTCATCACAGCAAACGTTTCGGAAAGACCGTTCGCCAGATTAACCCCACCCGAC
Top     PheAsnSerTrpArgIleHisHisSerLysArgPheGlyLysThrValArgGluIleAsnAlaThrArgP
         LeuIleProGlyAlaPheIleThrAlaAsnValSerGluArgArgPheAlaArgLeuT
        Leu . PheLeuAlaHisSerSerGlnGlnThrPheArgLysAspGlySerProAsp .
Bottom
        LysLeuGluGlnArgMet  .  . LeuLeuArgLysProPheValThrArgTrpIleLeu
         LysIleGlyProAlaAsnMetValAlaPheThrGluSerLeuArgAsnAlaLeuAsnVa
          . AsnArgAlaCysGluAspCysCysValAsnArgPheSerProGluGlySer . A
```

```
Position: 364
   370        380
CGAGCTGTATGC-GGAGT-GC
```

```
        86400     86600     86800     87000
8708-azr.abi(53>564)
3217fbjq.abi(34>555)
1403-atn.abi(58>684)
1859fbkc.abi(32>598)
1598-aul.abi(51>674)
4331-bae.abi(68>696)
2952-auc.abi(57>721)
0087fbfz.abi(34>686)
```

Fig. 1. The left side shows the 93.8-kb *SeqManII* project. Upper left is the Alignment Window with traces (four-color when viewed on color monitors) and consensus sequence aligned with six-frame translations. Dots in the translations are stop codons. Bottom left is the strategy view, showing where each sequence read covers the contig. The alignment view on the right shows a small project where *SeqManII*'s trace-quality-based consensus caller has correctly called an A in the consensus sequence when the basecalls in the corresponding column were T, T, and A (arrows). A majority-based consensus caller would incorrectly call the consensus a T at this position.

The **Strategy View (Fig. 1)** displays depth and orientation of sequence coverage throughout the contig. Thresholds for "complete" and "partial" coverage can be chosen by the user. In this *E. coli* project, complete was defined as fourfold coverage, with at least two reads in each direction. Different colors and thicknesses for the contig indicate where coverage is only a single read or



the Align-
consensus
ɔp codons.
the contig.
II's trace-
; sequence
(arrows).
ı T at this

ence cov-
coverage
efined as
ɜnt colors
le read or

on only one strand. In this example, a region around 87,000 nucleotides has full coverage but regions close by do not (**Fig. 1**). The **Strategy Viewer** simplifies decisions as to whether additional experiments are needed to complete the project.

2.7. Further Analysis

Once satisfied with a contig, any or all of the sequence may be selected as a query for *BLAST* searching over the internet using NCBI's *BLAST* server. Results are returned detailing public sequences with matches to the contig. This is a useful preliminary indication of the nature of the sequence, and if segments of the sequence are closely related to published ones, it can serve as an independent way to assess the assembly.

The contig may be exported as a single file in *DNASTAR*, *GenBank*, or *FASTA* formats. For sequences that do not match public data closely over their whole length, a typical next step would be gene discovery using *GeneQuest*.

3. Gene Discovery using *GeneQuest*

GeneQuest identifies a wide range of features in DNA sequences, and provides tools for annotation and visualization. Once characterized, the information may be submitted to public databases.

3.1. Sequence Entry

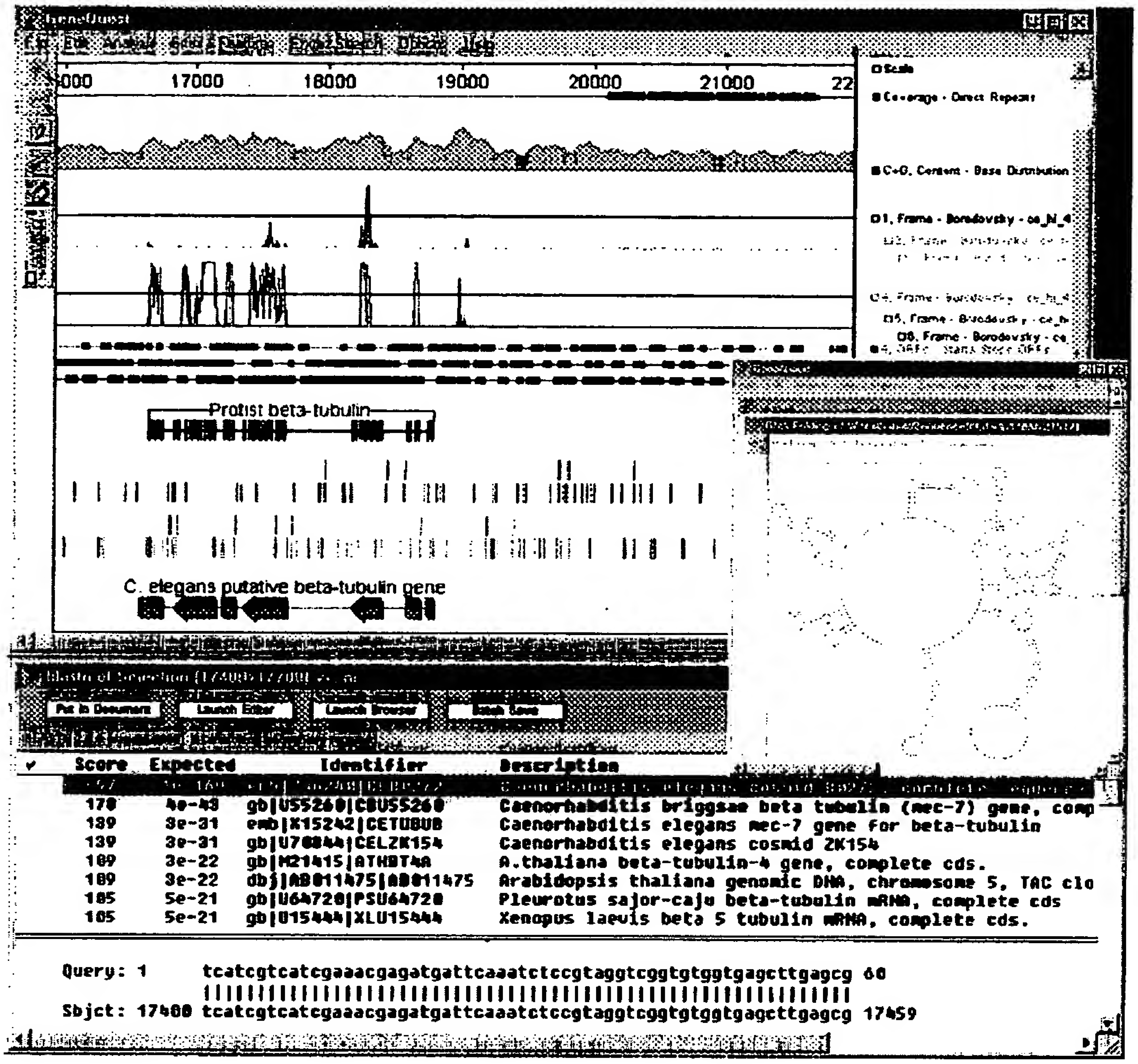
GeneQuest accepts sequence files in *DNASTAR, Genbank, ABI* and *SCF3* formats. Projects may be any size up to whole bacterial genomes. As an example, this section describes how gene discovery might proceed for a 28.67-kb cosmid clone of *Caenorhabditis elegans* imported from *GenBank* (accession no. Z46240).

3.2. Finding Coding Regions

3.2.1. Repeat Sequences and Base Distribution

Highly repetitive regions are unlikely to encode proteins. Thus finding repeats can exclude segments of DNA from the search for genes. *GeneQuest* identifies direct, dyad, and inverted repeat sequences. The occurrence of direct repeats around 20–21.7 kb of the *C. elegans* sequence (**Fig. 2**) suggests the absence of coding potential in this region.

Determining base content can assist in gene identification in organisms that have distinct G+C contents in coding vs. noncoding regions. For this example, a window of 50 nucleotides was used to average G+C content (**Fig. 2**). G+C content varies nonrandomly along the length of this sequence, with regions of higher G+C interspersed with stretches of lower G+C. One might focus first on the higher G+C regions for gene discovery.



```
✓    Score  Expected      Identifier                Description
     [illegible]
      170    4e-43    gb|U55260|CBU55260        Caenorhabditis briggsae beta tubulin (mec-7) gene, comp
      139    3e-31    emb|X15242|CETUBUB        Caenorhabditis elegans mec-7 gene for beta-tubulin
      139    3e-31    gb|U70844|CELZK154        Caenorhabditis elegans cosmid ZK154
      109    3e-22    gb|M21415|ATHBT4A         A.thaliana beta-tubulin-4 gene, complete cds.
      109    3e-22    dbj|AB011475|AB011475     Arabidopsis thaliana genomic DNA, chromosome 5, TAC clo
      105    5e-21    gb|U64720|PSU64720        Pleurotus sajor-caju beta-tubulin mRNA, complete cds
      105    5e-21    gb|015444|XLU15444        Xenopus laevis beta 5 tubulin mRNA, complete cds.

Query: 1      tcatcgtcatcgaaacgagatgattcaaatctccgtaggtcggtgtggtgagcttgagcg 60
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 17400  tcatcgtcatcgaaacgagatgattcaaatctccgtaggtcggtgtggtgagcttgagcg 17459
```

Fig. 2. The upper window shows *GeneQuest*'s main view, the "Assay Surface," which displays method results. Below the scale at top the first line shows direct repeats, and the graph below charts G+C content. The next plot is Borodovsky statistics, superimposed for frames 1–3, with little evidence of coding potential. The following bars indicate ORFs for frames 4–6, and immediately above them are the Borodovsky data for the same frames, consistent with the presence of a multi-exon gene. The *Gene Finder* result below shows that the protist β-tubulin coincides with the Borodovsky peaks. Viewing all these data in the context of the splice sites (next sets of hash marks) permits prediction of intron–exon boundaries. The authors of the sequence submission annotated a tubulin gene in the same location *GeneQuest* finds one. Bottom right shows predicted folding for RNA from the repeat sequence. Bottom left shows the results of a *BLAST* search using the ORF spanning coordinates 17,700–17,400 as query.

3.2.2. ORFs, Stops, and Starts

In prokaryotes, which lack introns, coding regions may be identified from sequence data by seeking long ORFs between start codons and stop codons, though this simple approach is rarely sufficient as a gene-finding tool. For eukaryotes, where introns are the norm and exons may comprise a minority of the sequence that makes up a gene, even genuine ORFs are typically much shorter than genes. **Figure 2** shows the ORFs and stops for the *C. elegans* sequence. *GeneQuest* displays plenty of candidate coding regions, but few ORFs are larger than 500 nucleotides long.

3.2.3. Coding Regions—Borodovsky Statistics

Statistical methods of identifying candidate coding regions offer considerable power, and are indispensable for finding coding regions in eukaryotes. Borodovsky's method ***(4)*** finds sequence patterns characteristic of coding regions. *C. elegans*-specific matrix files were used with this method to generate Borodovsky plots for all six reading frames (**Fig. 2**). There are compelling statistical indications of coding capacity for frames 4, 5, and 6 around sequence coordinates 19,000–16,500 nucleotides (**Fig. 2**). There is only marginal evidence for coding capacity in frames 1–3, and ≃2 kb on either side of this region there is little evidence at all. These results are consistent with the presence of a multi-exon gene in frames 4, 5, and 6.

3.2.4. Related Published Sequences—BLAST *searching*

For a fast indication of what a sequence might encode, a region with coding potential, such as the *C. elegans* segment 17,700–17,400 nucleotides, may be selected as query for a *blastn* search of *GenBank* over the Internet. In this case the sequence naturally found itself as the top match, as it was already published (**Fig. 2**). However, all of the top matches to the query are β-tubulins, and the matches are highly significant. This indicates that the segment used as query may be part of a *β-tubulin* gene.

3.2.5. Search for Specific Genes—Gene Finder

The coding region predicted in this example appears closely related to a known sequence. *GeneQuest*'s *Gene Finder* functions may be used to test explicitly for relatedness to β-tubulin. A file for a prototypical β-tubulin polypeptide (*GenBank* accession no. M58521) from a Protist was chosen as the protein to find. **Figure 2** shows where the regions of this tubulin match *C. elegans* codons. There is compelling overlap with the cluster of Borodovsky peaks, strongly indicating the presence of a *β-tubulin* gene.

3.2.6. Splice Sites—Statistical Patterns

GeneQuest provides statistical methods to locate intron–exon boundaries. In this example, *C. elegans*-specific matrix files were used to predict potential splice sites (**Fig. 2**).

The precise limits of the coding region and the intron–exon boundaries may be investigated by viewing the predicted splice sites in the context of the ORFs—ideally, the Borodovsky peak should coincide with an ORF in the same reading frame, bounded by donor and acceptor splice sites. If candidate splice sites are not found using the statistical pattern method, one can zoom in to magnify the *GeneQuest* display so that individual bases can be resolved and appropriate dinucleotides for splice sites may then be sought by eye.

Putting together all the information now gathered for this *C. elegans* example, the tubulin-coding region could be divided into seven exons (**Fig. 2**). As this example sequence was already published, one can examine the published annotations. Indeed, the authors of the *GenBank* submission had already annotated a putative 7-exon *β-tubulin* gene in the same location that *GeneQuest* predicts one (**Fig. 2**).

3.2.7. Other Discovery and Annotation Functions

In addition to the methods cited in the *C. elegans* example, *GeneQuest* can identify transcription factor binding sites, restriction sites, any pattern typed in by the user, codon usage for all or part of the sequence, and regions of DNA that are likely to bend *(5)*. *GeneQuest* can also simulate separation of restriction fragments by agarose gel electrophoresis, and predict how RNA corresponding to the selected DNA will fold (**Fig. 2**).

3.3. Text Searches to Find Related Sequences

GeneQuest provides access to the National Center for Biotechnology Information (NCBI) *Entrez* server, which processes text queries of sequence databases and returns the results over the internet. For example, to find other tubulins from *C. elegans*, one could build an *Entrez* query of the nucleic acid data for ["tubulin" in a text field] **AND** ["*Caenorhabditis*" in the organism field]. In mid 1998 this would match 68 database entries containing α, β- and γ-tubulins from *C. elegans*, tubulins from other members of the genus, tubulin tyrosine ligases from the genus, and other entries that refer to tubulins. However, the *Entrez* server does not support searches as sophisticated as the Boolean text/sequence searches that *GeneMan* provides (**Subheading 6.**).

3.4. Annotating and Displaying Features

GeneQuest provides tools for displaying and annotating features, whether they are found using *GeneQuest*, or in public data files imported into

ɔundaries.
t potential

.aries may
ɜxt of the
ı the same
late splice
ɔom in to
olved and
.
'. *elegans*
s (**Fig. 2**).
ɜ the pub-
ıd already
ꞔeneQuest

Quest can
n typed in
s of DNA
of restric-
NA corre-

ogy Infor-
ence data-
find other
ıcleic acid
organism
α, β- and
ıs, tubulin
ins. How-
the Bool-
).

s, whether
ɔrted into

GeneQuest. Once a region of interest is selected with the range selector tool, *GeneQuest* enters the sequence coordinates automatically into the annotation window when the **Features_New Feature** menu option is chosen. Information ranging from standard feature descriptions to free-form notes may be added. To annotate the *tubulin* gene in this example, the first exon was selected and annotated as a feature, then each of the other exons was selected and joined to the first exon. If conclusions change as to the boundaries of any features, the sequence coordinates may be adjusted in the annotation window, saving the effort of creating annotations anew.

Features may be displayed in a variety of forms, including graphs, boxes, arrows, bars, hash marks, and text, depending on the feature (**Fig. 2**). Colors and fill patterns may be selected from the *GeneQuest* tool kit, and graphic elements may be superimposed, rearranged, and juxtaposed by dragging them up and down. A useful approach is to specify, e.g., red, blue, and green for reading frames 1, 2, and 3, and use these colors consistently for frame-specific elements such as ORFs, stops, starts, and Borodovsky plots.

3.5. Creating Sets of Methods for Re-use

The range of analytical methods in *GeneQuest* is broad, and the scope for varying method parameters greatly multiplies the combinations of methods available for analyzing DNA sequences. Once a set of methods has been applied to a sequence with specific parameters, *GeneQuest* can save the **Method Outline,** which is analogous to a template used in word processing. The same methods and parameters can be quickly applied to another sequence simply by applying a saved Method Outline.

3.6. Further Analysis and Publication of the Results

To use external illustration functions, *GeneQuest*'s graphical views may be copied to the clipboard, then moved to other applications using the **Paste_Special** option and choosing to paste the picture (rather than the sequence). To edit the picture in *Microsoft Powerpoint*, for example, one may use the drawing tools to ungroup the pasted image. Each element—graphs, bars, arrows, labels, legends, and so on—may then be separately edited, resized, moved, or deleted.

GeneQuest saves projects as *GeneQuest* document files. For further analysis in other *Lasergene* applications, data may be saved as *DNASTAR* or *FASTA* files. For submission to public databases, the sequence and annotations may be saved as a *GenBank* flatfile.

4. Protein Structure Analysis with *Protean*

Protean works the same way *GeneQuest* does, but with polypeptides rather than DNA sequences. *Protean* accepts sequence files in *DNASTAR* format. For files in other formats, sequences may be converted to *DNASTAR* format using the *EditSeq* module (**Subheading 7.**).

4.1. Analyzing Sequences

Protean has over 20 analytical methods for predicting secondary structural and physicochemical properties of proteins. As with *GeneQuest* (**Subheading 3.**), most methods provide for customized graphical display of the results. For the example in this section, a human calmodulin protein was analyzed.

4.1.1. Predicting Alpha Helices, Beta Sheets, Coils, and Turns

Protean provides four methods for predicting secondary structures *(6–9)*. In this example, the Garnier-Robson method *(7)* was used to predict helices and turns. Calmodulins are well-documented proteins, so it is possible to view how *Protean*'s *in silico* predictions of helices and turns compare with reality—very well in this case (**Fig. 3**). For proteins that are not well characterized, use of multiple methods to analyze secondary structures is recommended.

4.1.2. Predicting Hydropathy and Amphiphilicity

Three methods are provided for predicting hydropathic character *(10–12)*. Hydrophobicity plots may also be predicted using the Eisenberg method *(13)*, which predicts amphipathic character. For the calmodulin example, the Kyte-Doolittle method *(11)* predicts that calmodulin is hydrophilic over most of its length (**Fig. 3**), and thus is unlikely to be embedded in membranes—again, a good match with reality.

4.1.3. Finding Motifs and Sequence Similarities

Searching the *PROSITE* database *(14)* for published motifs that match one or more segments of a protein, *Protean* located the four known "EF-hand" calcium-binding sites in calmodulin protein. They are located in the predicted turn regions (**Fig. 3**)—independent evidence of structural similarity for the four sites.

Protean provides *BLAST* search capability. As with *GeneQuest,* the results of a *BLAST* search may vary substantially depending on whether the whole sequence or a specific segment is chosen as the query. The ability to choose a specific segment of the protein sequence increases the probability of finding matches to a motif or structural element in unrelated proteins compared with a query based on the entire sequence.

4.1.4. Predicting Antigenic Regions, Surface Probability, and Flexibility

Identifying antigenic regions *in silico* can aid generation of useful antibodies greatly. *Protean* provides four methods for predicting antigenicity *(15–18)*. The Jameson-Wolf method *(16)* indicates five highly antigenic regions in this nat using

structural
ading 3.),
s. For the

;

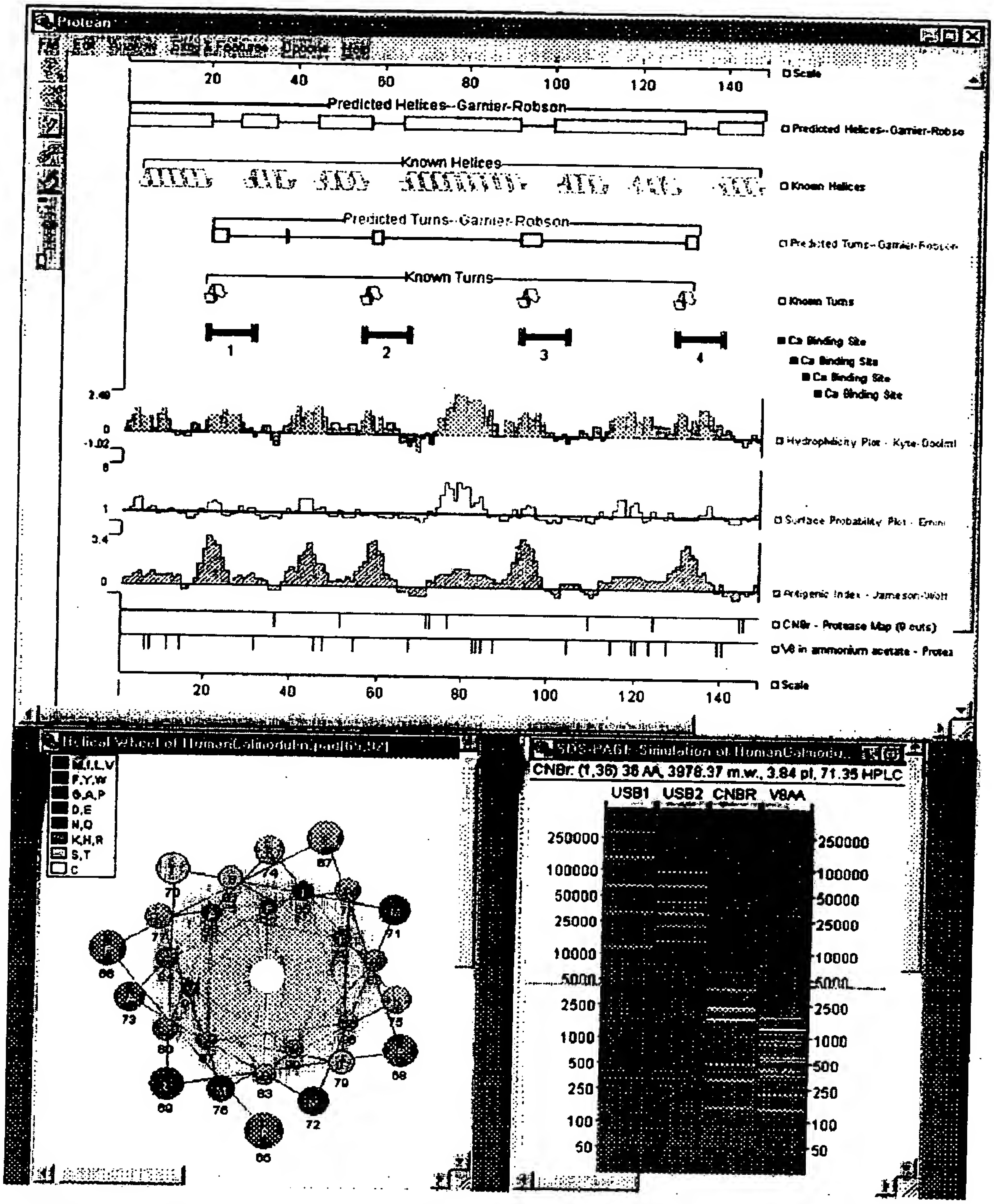
(6–9). In
:lices and
view how
ity—very
:d, use of

(10–12).
hod (13),
the Kyte-
ıost of its
—again, a

ıatch one
EF-hand"
licted turn
ɔur sites.
he results
he whole
choose a
ɔf finding
'ed with a

‹ibility

l antibod-
· (15–18).
ıns in this


Protean
File Edit Sequence Sites & Features Options Help
20
40
60
80
100
120
140
□ Scale
Predicted Helices--Garnier-Robson
□ Predicted Helices--Garnier-Robso
Known Helices
□ Known Helices
Predicted Turns--Garnier-Robson
□ Predicted Turns--Garnier-Robson
Known Turns
□ Known Turns
1
2
3
4
■ Ca Binding Site
■ Ca Binding Site
■ Ca Binding Site
■ Ca Binding Site
2.40
0
-1.02
□ Hydrophilicity Plot - Kyte-Doolittl
6
1
□ Surface Probability Plot - Emini
3.4
0
□ Antigenic Index - Jameson-Wolf
□ CNBr - Protease Map (9 cuts)
□ V8 in ammonium acetate - Protea
□ Scale
Helical Wheel of HumanCalmodulin.pad[65,92]
M,I,L,V
F,Y,W
G,A,P
D,E
N,Q
K,H,R
S,T
C
67
74
71
70
77
66
73
80
75
68
69
76
83
79
72
65
SDS-PAGE Simulation of HumanCalmodu...
USB1
USB2
CNBR
V8AA
250000
100000
50000
25000
10000
5000
2500
1000
500
250
100
50

CNBr: (1,36) 38 AA, 3978.37 m.w., 3.84 pI, 71.35 HPLC

Fig. 3. The upper window shows *Protean*'s main view. Bottom right is a simulation of electrophoretic separation of fragments of the sequence that would be generated by CNBR and trypsin (TRYT), compared with known molecular size standards (USB1, USB2, BRL); the characteristics of the selected band are displayed above the gel image. Bottom left displays a schematic view of the alpha-helical segment of the protein from residues 65–92, as seen by an imaginary observer looking down the center of the helix.

small protein (**Fig. 3**). Four coincide with the calcium-binding sites. Antibodies generated from these may cross-react with EF-hand calcium-binding sites in other types of proteins. The *Protean* analysis suggests that the antigenic region around residues 40–50 may be a better choice of antigen if reaction with other calcium-binding proteins is to be avoided.

Surface regions are often good antigens for antibodies for immunocytochemistry—antibodies generated may yield better experimental results if the epitope lies on the surface of the native protein. For this calmodulin protein, the Emni method *(19)* suggests that surface probability is moderate for the region around residues 40–50, indicating that this is an acceptable target region for immunocytochemical experiments. However, the Emni method also points to the hydrophilic region around residues 75–85 as a prominent surface region (**Fig. 3**). This region is predicted to have moderate antigenic character, and does not overlap any calcium-binding site. It may be a good alternative antigen to the 40–50 region.

The search for antigenic sites may be combined with the built-in *BLAST* search, to limit candidate antigenic segments to those that have suitable sequence specificity.

4.1.5. Finding Proteolytic Sites

Protean has a database of 24 proteolytic activities, including enzymes and chemical agents. A built-in editor allows addition or modification of activities. For calmodulin, cleavage sites for cyanogen bromide (CNBR) and V8 protease are shown (**Fig. 3**). As with *GeneQuest* (**Subheading 3.2.7.**), simulations of electrophoretic separation of proteolytic fragments can be displayed (**Fig. 3**).

4.1.6. Modeling Structures

Protean displays model structures for all or part of a sequence. Models include helical wheel (**Fig. 3**), helical net, beta net, linear space fill, and chemical formula. These models provide clues as to the three-dimensional structure of the gene product.

4.2. Creating Sets of Methods for Re-Use

To make repeated analyses with similar groups of methods more efficient, *Protean* can save and re-use "Method Outlines" just like *GeneQuest* (**Subheading 3.5.**).

4.3. Publication of the Results

As with *GeneQuest* (**Subheading 3.6.**), the graphical views in *Protean* may be altered using the built-in graphing tools, or copied to the clipboard for export to other applications for further modification.

ıntibodies
ıg sites in
nic region
with other

:ytochem-
he epitope
the Emni
on around
· immuno-
nts to the
ce region
acter, and
ve antigen

in *BLAST*
ɜ suitable

:ymes and
activities.
8 protease
ılations of
l (**Fig. 3**).

ɜ. Models
nd chemi-
l structure

: efficient,
eneQuest

ɔtean may
for export

5. Sequence Alignment/Phylogenetic Tree Building with *MegAlign*

MegAlign makes pairwise or multiple alignments of DNA or protein sequences, and generates graphical views of sequence similarities and differences, phylogenetic trees, and tables of the numerical data underlying the comparisons.

5.1. Sequence Entry from Files or Public Databases

MegAlign accepts sequence directly from *DNASTAR* files, and from ABI and SCF3 trace files. When both DNA and protein sequences are entered, *MegAlign* translates the DNA sequences and aligns all sequences as proteins unless set to backtranslate proteins for alignment as DNA.

If a protein is new to the investigator, sending the protein sequence as query to the NCBI *BLAST* server usually returns a list of similar sequences. Any of these may then be added directly into *MegAlign*.

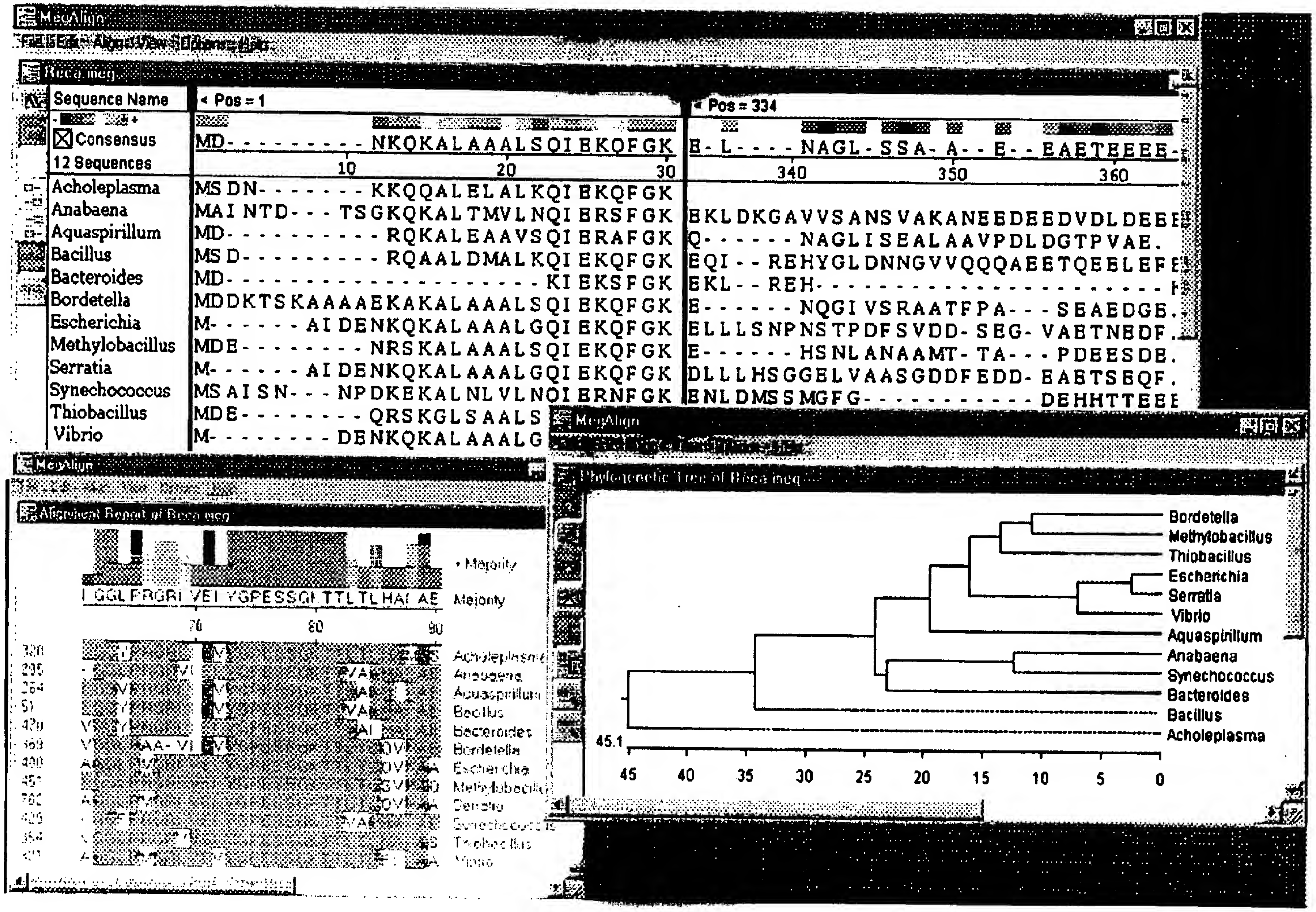
5.2. Aligning Multiple Sequences

Multiple alignments may be performed using *Jotun Hein (23)* and *ClustalV (24)* algorithms. The two methods typically yield slightly different results, both for the sequence alignment and for the phylogenetic tree generated. The ability to compare results using two methods is important, as it underscores the need for caution in interpreting the results.

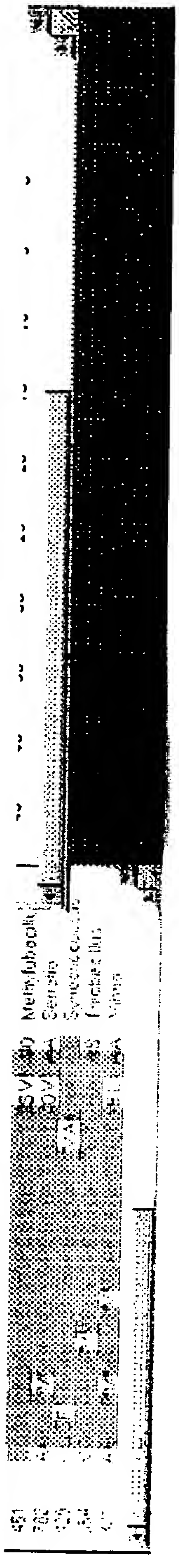
For the example in this section, a set of bacterial RecA sequences was aligned. The *MegAlign Worktable* (**Fig. 4**) is where the user edits and arranges the sequence names, and adjusts the segment of each sequence for alignment. Following alignment, the Worktable provides tools for making small adjustments to the alignment. Two panels display the left and right ends of the alignment, so that when adjustments to one region of the alignment are made, the effects on another region can be viewed immediately. The consensus sequence may be displayed above the alignment in the Worktable (**Fig. 4**).

A useful strategy is to perform an alignment on a set of complete sequences, and use the results to choose a common subsegment for final alignment. **Figure 4** shows a preliminary alignment of the RecA sequences. Most of the sequence regions match quite well, but the beginning and end show major variation that could make these regions unsuitable for inclusion in phylogenetic tree data. At this point, the sequences could be trimmed and realigned, using either the original or modified alignment parameters.

Users may experiment with realigning the sequences using different alignment parameters. A phylogenetic tree may be viewed at any time during the alignment process (**Fig. 4**). This is important, as minor editing of the alignment and changes to alignment parameters each may have a measurable impact on the phylogenetic trees built from the alignments. Iteratively changing align-



```
Sequence Name      < Pos = 1                                  < Pos = 334
☒ Consensus        MD- - - - - - - - - - NKQKALAAALSQIEKQFGK  E- L- - - - NAGL- SSA- A- - E- - EAETEEEE-
12 Sequences                      10         20         30        340          350          360
Acholeplasma       MS DN- - - - - - - KKQQALELALKQIEKQFGK
Anabaena           MAI NTD- - - TSGKQKALTMVLNQIERSFGK     EKLDKGAVVSANSVAKANEEDEEDVDLDEEE
Aquaspirillum      MD- - - - - - - - - - RQKALEAAVSQIERAFGK  Q- - - - - - NAGLISEALAAVPDLDGTPVAE.
Bacillus           MS D- - - - - - - - - RQAALDMALKQIEKQFGK  EQI- - REHYGLDNNGVVQQQAEETQEELEFE
Bacteroides        MD- - - - - - - - - - - - - - - - - KIEKSFGK  EKL- - REH- - - - - - - - - - - - - - - - - - - -
Bordetella         MDDKTSKAAAAEKAKALAAALSQIEKQFGK       E- - - - - - NQGIVSRAATFPA- - - SEAEDGE.
Escherichia        M- - - - - - AIDENKQKALAAALGQIEKQFGK   ELLLSNPNSTPDFSVDD- SEG- VAETNEDF.
Methylobacillus    MDE- - - - - - - - - NRSKALAAALSQIEKQFGK  E- - - - - - HSNLANAAMT- TA- - - PDEESDE.
Serratia           M- - - - - - AIDENKQKALAAALGQIEKQFGK   DLLLHSGGELVAASGDDFEDD- EAETSEQF.
Synechococcus      MSAISN- - - NPDKEKALNLVLNQIERNFGK      ENLDMSSMGFG- - - - - - - - - - - - DEHHTTEEE
Thiobacillus       MDE- - - - - - - - - QRSKGLSAALS
Vibrio             M- - - - - - - - - DENKQKALAAALG
```



ment parameters and viewing the resulting trees can distinguish phylogenetic placements that are robust to parameter changes vs those that are sensitive. The latter class would warrant more attention in refining the phylogeny.

5.3. Pairwise Alignments

Four algorithms are available for pairwise alignments: Martinez Needleman-Wunsch *(20)* and Wilbur-Lipman *(21)* for DNA alignments, and Lipman-Pearson *(22)* for protein alignments. The fourth method, dot plot, may be applied to both DNA and protein.

Pairwise alignments may be done on any two sequences selected in the Worktable window—there is no need to remove other sequences. Parameters may be adjusted for each pairwise alignment, and for each the result is displayed in its own window.

If part of a pairwise alignment seems unsatisfactory, that segment may be selected and realigned, either using the original alignment parameters or after changing them. This is particularly useful for optimizing longer alignments.

5.4. Viewing and Publishing Results

The "Alignment Report" view (**Fig. 4**) gives a customizable display of the same data shown in the Worktable. Adjustments may be made to the number of residues per line, the typeface, whether to display the sequences, and/or the consensus, whether to display a graphical representation of the similarity across the alignment, and how to display the individual differences and similarities—for example, identical residues may be shaded or hidden to emphasize residues that differ from the consensus.

As with other *Lasergene* applications, the Worktable view, alignment report, and phylogenetic tree may be copied to the clipboard and pasted into other applications (**Subheading 3.6.**) for further illustration. If additional computational analyses on the numerical data are desired, the sequence distance and residue substitution tables may be pasted directly into a *Microsoft Excel*® spreadsheet.

To use the data in other applications, they may be saved in formats suitable for use in the PAUP and GCG *Pileup* programs. The sequences in the *MegAlign*

Fig. 4. *(opposite page) MegAlign*'s *Worktable* (upper window) is where sequences are added and aligned, and where manual adjustments to the alignment may be made. The two panels show different segments of an alignment of 12 RecA-like proteins. This alignment took about five seconds using the *Clustal* method on a 200-MHz *PentiumPro* PC running *Windows NT4*. Bottom left shows the customizable Alignment View. The bar-chart displays the extent of agreement for the consensus residue in each column. Bottom right shows the phylogenetic tree.

project may also be exported as a set of individual *DNASTAR* sequence files. This is useful if the data were originally obtained as a *MegAlign* document from a colleague, or from the results of a *BLAST* search. One can then analyze any of the individual sequences in *GeneQuest* or *Protean* to elucidate what the consequences may be of any sequence variations found among the sequences examined in *MegAlign*.

6. Finding Public Data using *GeneMan*

GeneMan searches public data that are stored locally on CD-ROMs. Databases include *GenBank/EMBL*, *GBTrans*, and *PIR/NBRF*. Users can search the multiple CD-ROMs directly, or load the data onto a local hard drive and search from a single location. The latter option is quite feasible since a 17-gb hard drive can be purchased in 1998 for approx $400.

6.1. Sequence Similarity

GeneMan searches public data based on sequence similarity for either protein or DNA, using a modified *FASTA* algorithm *(25)*. *GeneMan* formulates the query from a sequence file, and provides options to change the sequence coordinates to use as query, the k-tuple (unit of comparison), the window size over which similarity is calculated, the percent similarity required for a match, and the penalty for introducing a gap between the query and the database match when the two are aligned (**Fig. 5**). For DNA searches, the "rapid screen" option screens the database before the *FASTA*-based search, reducing the number of sequences that FASTA subsequently searches dramatically, thereby accelerating the search.

Display of search results is initially as one-line summaries. Users may choose to expand the view to include more or all of the information available. Additional displays include a plot of the frequency of database hits vs the percent similarity, and the alignment between the query and the database hit (**Fig. 5**).

6.2. Consensus Sequences

GeneMan's consensus search function accepts a consensus sequence query up to 256 characters, and permits adjustment of the percent similarity for a match. Searching for a consensus is a powerful way to find known motifs. The syntax for defining the consensus is based on the conventions for *Prosite (26)*. These conventions support IUB codes, and for any sequence position allow explicit specification of alternative residues, excluded residues, specific distances between residues, and whether a pattern must be located at the amino or carboxyl terminus.

An example of a consensus sequence is the tubulin GTP-binding consensus [SAG]-G-G-T-G-[SA]-G, which specifies S or A or G in position 1, followed by GGTG, followed by S or A, followed by G (all one-letter amino acid codes).

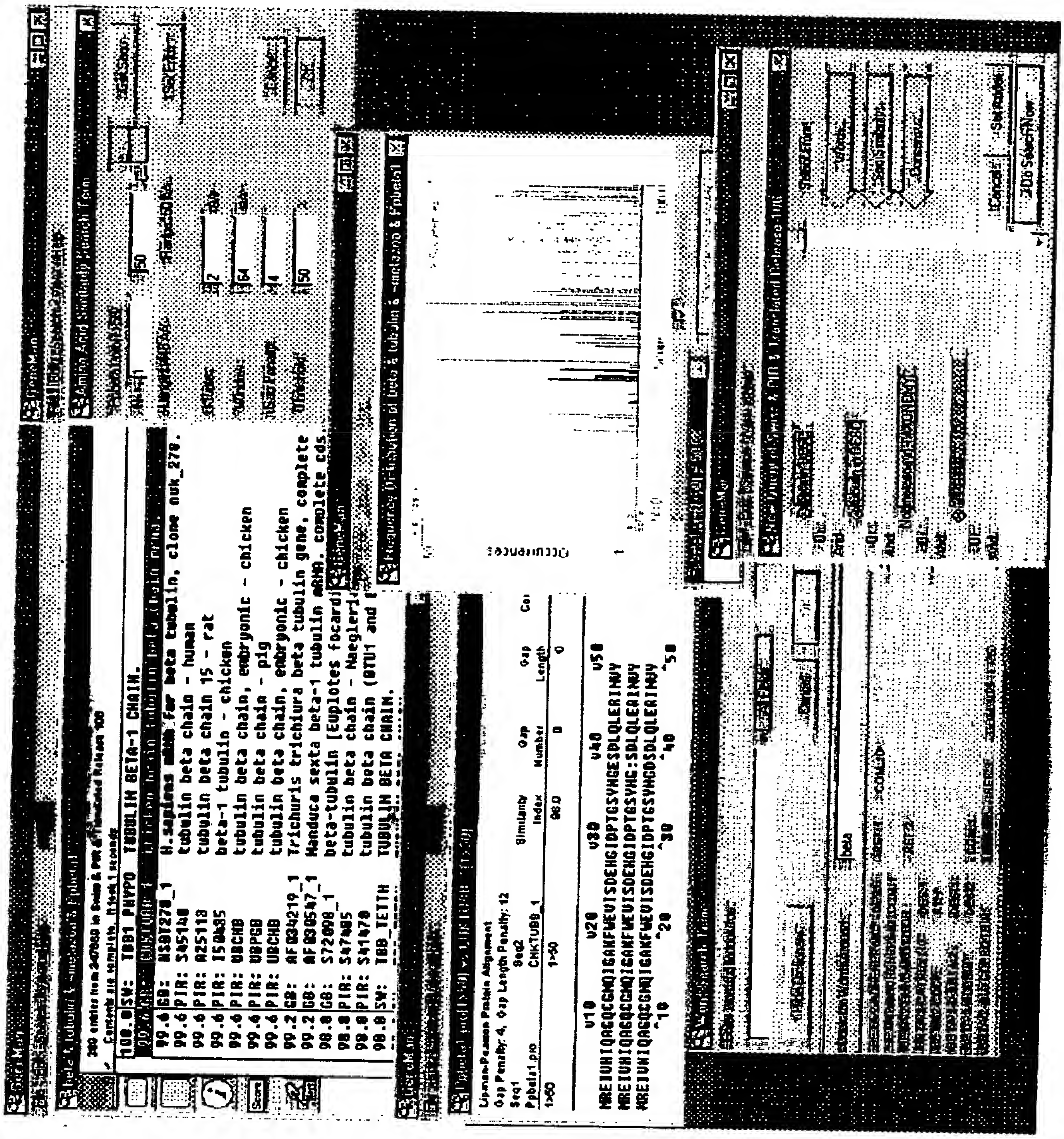
ience files.
document
en analyze
te what the
sequences

)Ms. Data-
ı search the
and search
17-gb hard

either pro-
mulates the
ience coor-
w size over
tch, and the
:h when the
ion screens
: sequences
the search.
Users may
ı available.
the percent
it (**Fig. 5**).

ence query
larity for a
notifs. The
rosite (26).
ition allow
ɔecific dis-
ie amino or

: consensus
l, followed
.cid codes).



```
100.0 SW:  TBB1_PHYPO   TUBULIN BETA-1 CHAIN.
99.6 GB:   HSBT278_1    H.sapiens mRNA for beta tubulin, clone nuk_278.
99.6 PIR:  S45140       tubulin beta chain - human
99.6 PIR:  A25118       tubulin beta chain 15 - rat
99.6 PIR:  I50485       beta-1 tubulin - chicken
99.6 PIR:  UBCHB        tubulin beta chain, embryonic - chicken
99.4 PIR:  UBPGB        tubulin beta chain - pig
99.4 PIR:  UBCHB        tubulin beta chain, embryonic - chicken
99.2 GB:   AF034219_1   Trichuris trichiura beta tubulin gene, complete
99.2 GB:   AF030547_1   Manduca sexta beta-1 tubulin mRNA, complete cds.
98.8 GB:   S72098_1     beta-tubulin [Euplotes focardi
98.8 PIR:  S47485       tubulin beta chain - Naegleri
98.8 PIR:  S41470       tubulin beta chain (BTU1 and
98.8 SW:   TBB_TETTH    TUBULIN BETA CHAIN.
```

```
Lipman-Pearson Protein Alignment
Gap Penalty: 4, Gap Length Penalty: 12
Seq1         Seq2         Similarity   Gap      Gap
Pbbeta1.pro  CHKTUBB_1    Index        Number   Length
1>50         1>50         98.0         0        0

         v10       v20       v30       v40       v50
MREIVHIQAGQCGNQIGAKFWEVISDEHGIDPTGSYHGESDLQLERINVY
MREIVHIQAGQCGNQIGAKFWEVISDEHGIDPTGSYHG:SDLQLERINVY
MREIVHIQAGQCGNQIGAKFWEVISDEHGIDPTGSYHGDSDLQLERINVY
         ^10       ^20       ^30       ^40       ^50
```

Fig. 5. Top left shows part of the match list for a sequence similarity search of the protein databases with a β-tubulin query. Immediately below is the alignment for the best match—alignments for any match may be displayed by selecting the match result with the mouse. Top and bottom right are dialogs for setting up queries and for adjusting parameters for sequence similarity terms. Bottom left is the word search dialog and associated dictionary listing the indexed words. Middle right shows a summary plot of the number of database matches with various levels of similarity to the sequence query.

6.3. Text Words

GeneMan provides tools to build a comprehensive text query before initiating a search, or to begin with a simple query and further query the results of the first search. In the latter case, the results for each stepwise query are kept in separate windows, allowing one to step back and perform multiple distinct searches on a prior set of results.

For each text word, a single field in the database may be specified, or all the fields may be searched. To check whether a word is indexed in the dictionary, the full dictionary may be viewed and any word selected from it (**Fig. 5**). Searching for multiple words, one can use Boolean operators AND, OR, NOT in combination with the field limitations. This allows quite specific queries to be built, increasing the focus on the data sought.

6.4. Complex Boolean Queries and Data Subsets

GeneMan can combine sequence similarity searching, text searching, and consensus searching in a single, comprehensive Boolean query. For example, imagine a user has found a potential GTP-binding site in a sequence. The next step might be identification of more gene products that have GTP-binding sites. However, the user knows this sequence is not a tubulin, and in searching for sequences encoding GTP-binding sites, wants to avoid perusing the thousand or so tubulin sequences in the public databases. One way to achieve this would be to combine with a text query for GTP binding the NOT operator for the tubulin GTP-binding consensus sequence [SAG]-G-G-T-G-[SA]-G.

Care is required in building the queries and interpreting the search results. As with any database search tool, *GeneMan* will not correct errors or omissions in the source data. Despite the best efforts of custodians, many entries in public databases contain errors, and annotations are not completely consistent. For example, to find all mammalian β-tubulin sequences, one could logically specify the text words "tubulin" **AND** "beta" in the **Definition** field, and the text word "mammal" in any field. This would find most mammalian β-tubulin sequence entries, but not those few β-tubulin sequence entries that do not have the word "tubulin" in the **Definition** field. However, searching for sequence similarity to β-tubulin, **AND** the text word "mammal" in any field, would find mammalian β-tubulins even if they lack "tubulin" in the **Definition** field, as β-tubulin sequences are highly conserved.

Similarly, imagine trying to formulate a query that eliminates Metazoan sequences from the hit list. Logically, this could be accomplished by choosing the Boolean operator **NOT** for the text query "Metazoa" restricted to the database field **Source.** But this query will produce a list of database entries that includes some sequences from the Metazoan *Drosophila,* simply because the

ore initiat-
sults of the
ıre kept in
ıle distinct

l, or all the
dictionary,
t (**Fig. 5**).
OR, NOT
queries to

ching, and
r example,
:. The next
ıding sites.
ırching for
e thousand
this would
tor for the
.
ch results.
s or omis-
y entries in
consistent.
d logically
ld, and the
ı β-tubulin
lo not have
r sequence
would find
n field, as

Metazoan
y choosing
o the data-
entries that
ecause the

term "Metazoa" is not in the **Source** field in some *Drosophila* database entries. *GeneMan*'s ability to combine sequence similarity searches with text searches provides the power to find relevant sequences even when database entries contain errors or omissions.

7. Primer Design, Restriction Maps, and Sequence Editing

The *Lasergene* system includes three more applications. These are described briefly.

PrimerSelect provides tools for design and analysis of oligonucleotides, including primers for PCR, sequencing, probe hybridization, and transcription. Using DNA, RNA, or backtranslated proteins as templates, *PrimerSelect* details thermodynamic properties for annealing reactions, identifies all possible primers, and ranks them in order of suitability for specified conditions. *PrimerSelect* also highlights potential pitfalls in both standard and multiplex PCR experiments.

MapDraw generates restriction maps and displays sites, translations, and features of sequences in six different formats, including circular and linear maps. The sequence entered may be as small as an oligonucleotide or as large as the largest BAC insert. From the database of nearly 500 restriction sites, any subset may be selected for mapping. Sets of restriction sites may be combined using Boolean operators, providing powerful site selection tools to assist cloning strategies.

EditSeq is provided with every *Lasergene* system to facilitate work on nucleic acid and protein sequences of all sizes from a variety of formats, including *GeneMan, GenBank, FASTA,* text, ABI, ALF, *Staden*, clipboard, GCG, *MacVector*™, and the efficient Lasergene sequence file format *DNASTAR* established in 1982. In addition, sequences may be obtained by accession number from NCBI's databases over the internet, and sequences related to a sequence in *EditSeq* may be identified using the integrated *BLAST* search tool. *EditSeq* provides basic analytical tools, including editing, reverse complementing, translation, back translation, ORF identification, and simple annotation.

8. Summary

Lasergene's eight modules provide tools that enable users to accomplish each step of sequence analysis, from trimming and assembly of sequence data, to gene discovery, annotation, gene product analysis, sequence similarity searches, sequence alignment, phylogenetic analysis, oligonucleotide primer design, cloning strategies, and publication of the results. The *Lasergene* software suite provides the functions and customization tools needed so that users can perform analyses the software writers never imagined.

Acknowledgments

I thank my colleagues Carolyn Allex and Sharon Savage for comments on the manuscript and Jeff Briganti for his patient explanations of all the myriad *Lasergene* features.

References

1. Blattner, F. R., Plunket III, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B., and Shao, Y. (1997) The complete genome sequence of *Escherichia coli* K12. *Science* **277,** 1453–1462.
2. Burland, V., Daniels, D. L., Plunkett III, G., and Blattner, F. R. (1993) Genome sequencing on both strands: the Janus strategy. *Nucleic Acids Res.* **21,** 3385–3390.
3. Allex, C. F., Baldwin, S. F., Shavlik, J. W., and Blattner, F. R. (1997) Increasing consensus accuracy in DNA fragment assemblies by incorporating fluorescent trace representations, in *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology* (Gaasterland, T., Karp, P., Karplus, K. Ouzounis, C., Sander, C., Valencia, A.), AAAI Press, Menlo Park, pp. 3–14.
4. Borodovsky, M. and McIninch, J. (1993) GeneMark: parallel gene recognition for both DNA strands. *Comp. Chem.* **17,** 123–133.
5. Trifonov, E. N. and Sussman, J. L. (1980) The pitch of chromatin DNA is reflected in its nucleotide sequence. *Proc. Natl. Acad. Sci. USA* **77,** 3816–3820.
6. Chou, P. Y. (1990) Prediction of protein structural classes from amino acid composition, in *Prediction of Protein Structure and the Principles of Protein Conformation* (Fasman, G. D., ed.), Plenum, New York, NY, pp. 549–586.
7. Garnier, J., Osguthorpe, D. J., and Robson, B. (1978) Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* **120,** 97–120.
8. Deléage, G. and Roux, B. (1987) An algorithm for protein secondary structure prediction based on class prediction. *Protein Eng.* **1,** 289–294.
9. Parry, D. A. (1982) Coiled coils in α-helix-containing proteins: analysis of the residue types in the heptad repeat and the use of these data in the prediction of coiled coils in other proteins. *Biosci. Rep.* **2,** 1017–1024.
10. Engelman, D. M., Steitz, T. A., and Goldman, A. (1986) Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu. Rev. Biophys. Biophys. Chem* **15,** 321–54.
11. Kyte, J. and Doolittle, R. F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157,** 105–132.
12. Hopp, T. P. and Woods, K.R. (1981) Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl. Acad. Sci. USA* **78,** 3824–3828.
13. Eisenberg, D., Weiss, R. M., and Terwilliger, T. C. (1984) The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc. Natl. Acad. Sci. USA* **81,** 140–144.

ıments on
he myriad

. Riley, M.,
:, J., Davis,
d Shao, Y.
:ience **277,**

3) Genome
ı385–3390.
Increasing
fluorescent
ference on
Karplus, K.
. 3–14.
›gnition for

is reflected

› acid com-
ein Confor-

ιe accuracy
tructure of

y structure

lysis of the
ediction of

g nonpolar
Annu. Rev.

the hydro-

eterminants
:8.
ydrophobic
d. Sci. USA

14. Bairoch, A., Bucher, P., and Hofmann, K. (1997) The PROSITE database, its status in 1997. *Nucleic Acids Res.* **25,** 217–221.
15. Margalit, H., Spouge, J. L., Cornette, J. L., Cease, K. B., Delisi, C., and Berzofsky, J. A. (1987) Prediction of immunodominant helper T cell antigenic sites from the primary sequence. *J. Immunol.* **138,** 2213–2229.
16. Jameson, B. A. and Wolf, H. (1988) The antigenic index: a novel algorithm for predicting antigenic determinants. *Comp. Appl. Biosci.* (now *Bioinformatics*) **4,** 181–186.
17. Sette, A., Buus, S., Appella, E., Smith, J. A., Chesnut, R., Miles, C., Colon, S. M., and Grey, H. M. (1989) Prediction of major histocompatibility complex binding regions of protein antigens by sequence pattern analysis. *Proc. Natl. Acad. Sci. USA* **86,** 3296–3300.
18. Rothbard, J. B. and Taylor, W. R. (1988) A sequence pattern common to T cell epitopes. *EMBO J.* **7,** 93–100.
19. Emini, E. A., Hughes, J., Perlow, D., and Boger, J. (1985) Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide. *J. Virol.* **55,** 836–839.
20. Martinez, H. M. (1983) An efficient method for finding repeats in molecular sequences. *Nucleic Acids Res.* **11,** 4629–4634.
21. Wilbur, W. J. and Lipman, D. J. (1983) Rapid similarity searches of nucleic acid and protein data banks. *Proc. Natl. Acad. Sci. USA* **80,** 726–730.
22. Lipman, D. J. and Pearson, W. R. (1985) Rapid and sensitive protein similarity searches. *Science* **227,** 1435–1441.
23. Hein, J. (1990) Unified approach to alignment and phylogenies. *Meth. Enzymol.* **183,** 626–645.
24. Higgins, D. G. and Sharp, P. M. (1989) Fast and sensitive multiple sequence alignments on a microcomputer. *Comp. Appl. Biosci.* (now *Bioinformatics*) **5,** 151–153.
25. Pearson, W. R. (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. *Meth. Enzymol.* **183,** 63–98.
26. Bucher, P. and Bairoch, A. (1994) A generalized profile syntax for biomolecular sequence motifs and its function in automatic sequence interpretation, in *Proceedings 2nd International Conference on Intelligent Systems for Molecular Biology* (Altman R., Brutlag D., Karp P., Lathrop R., Searls D., eds.), AAAI Press, Menlo Park, CA, pp. 53–61.